MINISTÈRE
DE L'ÉDUCATION
NATIONALE,
DE LA JEUNESSE
ET DES SPORTS

Liberté
Égalité
Fraternité

VOIE TECHNOLOGIQUE

Série STMG : Sciences et technologies du management et de la gestion

2DE | 1RE | TLE

Management, sciences de gestion et numérique

ENSEIGNEMENT DE SPECIALITE

DATA LAKE : DÉFINITION, AVANTAGES ET INCONVÉNIENTS POUR L'ENTREPRISE[1]

Article du site LeBigData.fr, dédié au big data et cloud computing
Par Bastien L., publié le 10 juillet 2017

Les *Data Lakes*, ou lacs de données, sont de plus en plus utilisés par les entreprises pour le stockage de données. Découvrez la définition du Data Lake, ses avantages, ses inconvénients, et ses différences avec le Data Warehouse.

Un Data Lake est un référentiel de données permettant de stocker une très large quantité de données brutes dans le format natif pour une durée indéterminée. Cette méthode de stockage permet de faciliter la cohabitation entre les différents schémas et formes structurelles de données, généralement des blobs d'objets ou des fichiers.

Au sein d'un seul Data Lake, toutes les données de l'entreprise sont stockées. Les données brutes, y compris les copies des données système source, côtoient les données transformées. Ces données sont ensuite utilisées pour établir des rapports, pour visualiser les données, pour l'analyse de données ou pour le Machine Learning.

Data Lake : un lac de données utile

Le Data Lake regroupe les données structurées en provenance de bases de données relationnelles en couloir ou en colonne, les données semi-structurées telles que les CSV, les logs, les XML, les JSON, et les données non structurées telles que les emails, les documents et les PDF. On y trouve même des données binaires telles que des images, des fichiers audio ou des vidéos.

1. Article consulté le 09/12/2020 : https://www.lebigdata.fr/data-lake-definition

Le terme de Data Lake fut conceptualisé pour la première fois par James Dixon, CTO de Penthao, pour établir un parallèle avec le *Data Mart*. Le *Data Mart* est un référentiel de données plus petit regroupant des attributs intéressants extraits de données brutes. Selon lui, le *Data Mart* présentait plusieurs problèmes, et le Data Lake se présentait comme la solution optimale. Le principal problème du *Mart* était le phénomène du silo d'informations. Comme l'a démontré une étude de PricewathouseCoopers, le Data Lake peut mettre de mettre fin à ce problème, car les entreprises peuvent s'en servir pour extraire leurs données et les entreposer au sein d'un seul référentiel Hadoop.

Selon Dixon, si l'on considère un *Data Mart* comme un magasin d'eau en bouteille, empaquetée pour une consommation facile, le Data Lake quant à lui est une grande source d'eau à l'état naturel. Les différents utilisateurs peuvent venir examiner ce lac, se plonger dedans, ou en extraire des échantillons.

Exemples de Data Lake

En guise d'exemple de Data Lake, on peut citer le système de fichiers distribué Apache Hadoop. La première version d'Hadoop 1.0 avait des capacités limitées en termes de traitement de données Map Reduce. Pour interagir avec ce Data Lake, il était nécessaire de maîtriser Java, Map Reduce, et des outils de haut niveau comme Pig et Hive. L'arrivée de Hadoop 2.0, de YARN pour la gestion de ressources, et de nouveaux paradigmes de traitement comme le streaming, ont permis de surmonter ces limites.

Beaucoup d'entreprises utilisent également des services de stockage cloud comme Amazon S3. De plus en plus d'institutions scolaires s'intéressent aux Data Lake. Par exemple, la Cardiff University a mis en place le projet Personal DataLake, afin de créer un nouveau type de Data Lake permettant de gérer le Big Data d'utilisateurs individuels en leur fournissant un point centralisé de collecte, d'organisation et de partage de données personnelles.

Enfin, les entreprises qui ont recours à l'Internet des Objets sont très friandes du modèle Data Lake. En effet, il faut pouvoir rassembler les données en provenance de centaines, voire de millions de capteurs et de les corréler. Cette infrastructure est par exemple au cœur du fonctionnement des compteurs connecté Linky en cours d'installation dans toute la France par Enedis. Ces derniers relèvent différents types d'informations sur la consommation, la puissance allouée, les défauts de sécurité et facilitent l'intervention des équipes de maintenance.

Quels sont les inconvénients d'un Data Lake ?

Les avancées technologiques sont souvent perçues comme un mouvement inévitable vers l'avant, une fatalité pour l'industrie de la tech. Toutefois, il est incorrect de penser que la nouveauté est toujours synonyme d'amélioration. Les nouvelles technologies apportent bien souvent leur lot de défis. Certains sont prédictibles, d'autres moins. Ces challenges minimisent les bénéfices offerts par l'innovation.

Les *Data Lakes* illustrent à merveille ce phénomène. Les entreprises ont adopté les *Data Lakes* très rapidement, en les percevant comme un complément voire un remplacement des leurs *Data Marts* et de leurs Data Warehouses. Or, ces entreprises ont fait abstraction des limites et des défauts des *Data Lakes*.

En effet, les *Data Lakes* peuvent créer davantage de problèmes qu'ils n'en résolvent. Selon Adam Wray, CEO et Président de Basho, les *Data Lakes* sont tout simplement « maléfiques ». Il est préférable pour une entreprise de percevoir leurs données à travers un prisme de chaîne logistique doté d'un début, d'un milieu et d'une fin. Ces données doivent être collectées, trouvées, explorées et transformées en suivant un plan organisé. Cette approche permet de maximiser la valeur extraite des données.

Or, les *Data Lakes* peuvent totalement ruiner cette tactique. Les lacs de données permettent de stocker n'importe quel format sans limite de quantité, ce qui conduit à de nombreux problèmes et empêche d'extraire de la valeur des données. Sans capacité à catégoriser ou établir une hiérarchie entre les données, le désordre s'installe rapidement.

Les *Data Lakes* sont massivement adoptés car il semble évident qu'il est impossible de tirer profit de données qui ne sont pas à portée de main. Cependant, on ne tire pas toujours automatiquement de valeur des données à portée de main. Les *Data Lakes* ne permettent pas d'établir de priorité entre les données et de définir comment elles seront utilisées. En résulte un véritable musée regroupant d'innombrables œuvres d'art, mais dépourvu du regard d'un curateur capable de déterminer lesquelles méritent d'être exposées.

Pour Wray, la principale raison pour laquelle les *Data Lakes* sont maléfiques est qu'ils sont dépourvus de règles, incroyablement coûteux, et que la valeur qui peut en être extraite est minime comparée à ce qui est promis. Ce constat peut sembler évident et inévitable, face à l'immense quantité de données disponibles à notre époque. Cependant, pour Wray, ce phénomène est connecté à d'autres problèmes liés aux *Data Lakes*. Pour éviter ces problèmes, il estime qu'il est important que les entreprises se posent trois questions : les données sont-elles actuelles, où sont-elles, et quels sont les risques encourus en les stockant ?

Les données doivent être exploitables au moment opportun

Les données ne sont utiles que si elles peuvent permettre de prendre de bonnes décisions au moment opportun. Si une entreprise souhaite analyser des données et qu'elle doit passer beaucoup de temps à les trouver dans le Data Lake et à les préparer, l'efficience s'en trouve fortement réduite.

Concrètement, le data lake ne permet pas aux entreprises de prioriser leurs données. De fait, il est plus difficile de percevoir en quoi ces données peuvent être utiles. Pour Wray, le véritable objectif des données devrait être différent. Les entreprises devraient se demander comment fournir ces données sous une forme épurée, simplifiée, afin qu'un maximum de personnes puisse y accéder et agir. Or, les *Data Lakes* n'aident pas à répondre à cette question.

Dans la plupart des cas, les données devraient être synthétisées ou utilisées avant d'être stockées dans le Data Lake. Par exemple, les défenseurs du Data Lake affirment souvent que toutes les données en provenance des capteurs IoT devraient être stockées sur le Data Lake. Toutefois, si l'on prend l'exemple d'un capteur de température équipé à une turbine, si la température atteint un certain seuil, des mesures doivent être prises pour l'éteindre ou quelqu'un doit être envoyé pour la réparer. Il s'agit donc d'une information opportune. Même sur le long terme, si le

capteur transmet la température toutes les 15 secondes, il est inutile de stocker toutes les données du capteur s'il transmet la température toutes les 15 secondes. Les données peuvent être synthétisées tant que la température est stable, sans pour autant perdre la moindre information utile. Plutôt que de stocker tout et n'importe quoi, il est plus judicieux de prendre des décisions en se basant sur les données nécessaires.

Un problème de latence à prendre en compte

Dans un monde de plus en plus mondialisé, où il est possible d'accéder aux données depuis n'importe où dans le monde, on pourrait penser que la localisation des données n'importe plus. Toutefois, comme le souligne Wray, la latence dépend de la localisation des données. Cette latence peut nuire à l'opportunité des données.

Si les *Data Lakes* sont physiquement très éloignés, il peut être très long de récupérer une donnée spécifique. Or, les données sont utilisées trop souvent pour négliger ce phénomène. La latence peut ralentir l'entreprise dans son ensemble. De plus, la localisation du Data Lake peut également réduire la sécurité des données.

Avec les *Data Lakes*, les entreprises pensent souvent qu'elles peuvent se contenter de stocker toutes leurs données de n'importe quelle façon, sans tenir compte de leur provenance. Toutefois, un tel comportement expose l'entreprise à de nombreux risques, notamment à l'égard des lois.

Un risque pour la confidentialité des données

Comme le souligne Wray, les lois et les régulations relatives à la confidentialité des données diffèrent dans tous les pays. Il n'est donc pas possible de transposer un modèle d'un pays à l'autre comme si de rien n'était. Ce constat peut sembler évident, mais de nombreuses entreprises n'en tiennent pas compte. Le manque de hiérarchisation des données au sein des *Data Lakes* peut augmenter la difficulté à s'adapter aux régulations. Il est possible que les entreprises ne connaissent pas toutes les données qu'elle collecte, leur provenance, et les risques auxquels elles les exposent.

Les fuites de données sensibles sur les clients, tels que des informations financières ou d'e-mails privés, sont fréquentes. C'est pourquoi la protection des consommateurs est essentielle dans tous les business. L'absence de contraintes des *Data Lakes* peut conduire les entreprises à placer des données à risque à un endroit mal sécurisé. Près de 200 millions de votants américains en ont fait les frais il y a peu après qu'un data lake soit rendu disponible sur un Cloud public. Il s'agit de la fuite de données la plus importante de l'Histoire.

C'est la raison pour laquelle Wray considère que les *Data Lakes* doivent être maniés avec précaution. Il considère que ces lacs de données sont maléfiques, car ils se présentent de façon métaphorique comme un buffet de nourriture que l'on pourrait comparer à un repas dans un bon restaurant. La quantité est trop importante, et la qualité est médiocre. Selon lui, environ 95 % des données collectées et agrégées au sein du Data Lake seront jetées. Ainsi, beaucoup d'entreprises utilisent Hadoop et les *Data Lakes* en pensant que si le service est complexe et coûteux, il sera forcément de haute qualité. Il s'agit d'une idée absurde.

En réalité, les *Data Lakes* créent une complexité, ajoutent des charges, et sèment la confusion au sein de l'entreprise sans pour autant fournir une grande valeur. De nombreuses entreprises pensent stocker des données maintenant pour en extraire une valeur ultérieurement, mais n'y parviennent jamais car la difficulté pour dégager cette valeur est trop extrême. Ceci peut conduire à de véritables désastres sur le long terme.

Les chefs d'entreprise devraient cesser de stocker tout et n'importe quoi sur leurs *Data Lakes* et se focaliser sur des projets individuels, et sur les outils analytiques nécessaires pour fournir une grande valeur en assemblant uniquement les données nécessaires. Ce processus permet de créer des données actionnables pour résoudre les problèmes rencontrés par l'entreprise à l'heure actuelle. Après avoir complété une série de problèmes individuels, il est possible de chercher quelles données sont utilisées plus souvent et quelles données sont une priorité. Il est ensuite possible de créer un référentiel plus efficient et plus efficace.

Le Data Lake est un défi pour l'entreprise

Selon David Needle, les *Data Lakes* sont l'une de façons les plus controversées de gérer le Big Data. De même, PricewaterhouseCoopers tempérer son enthousiasme en précisant que toutes les initiatives Data Lake ne sont pas nécessairement des succès.

D'après Sean Martin, CTO de Cambridge Semantics, de nombreuses entreprises créent des cimetières Big Data, dans lesquels ils entreposent tous leurs fichiers Hadoop et espèrent en tirer quelque chose par un heureux hasard. Par la suite, ils oublient et perdre la trace des données ainsi stockées.

Ainsi, le principal défi n'est pas de créer un Data Lake, mais de tirer profit des opportunités qu'il représente. Les entreprises ayant développé des *Data Lakes* avec succès l'améliorent progressivement, à mesure qu'elles se demandent quelles données et métadonnées sont importantes pour leur activité

Différences entre Data Lake et Data Warehouse

Tandis qu'une Data Warehouse permet d'entreposer des données dans des fichiers ou des dossiers, un Data Lake repose sur une architecture de type flat. Chaque élément de donnée dans un Lake se voit assigner un identifiant unique, et tagué à l'aide d'un ensemble étendu de mots-clés de métadonnées. Si l'entreprise se pose une question, elle peut effectuer une requête pour chercher des données pertinentes au sein du Data Lake. Par la suite, l'ensemble de données délivré en réponse à cette requête peut être analysé pour répondre à la question.

Le terme Data Lake est bien souvent associé au stockage d'objet orienté vers Hadoop. Dans un tel scénario, les données d'une entreprise sont tout d'abord chargées sur la plateforme Hadoop, puis les outils de business analytiques et de *data mining* sont utilisés sur les données au sein des nœuds de cluster Hadoop où elles résident.

Tout comme le Big Data, le terme de Data Lake est parfois utilisé comme un simple label marketing pour un produit supportant Hadoop. De même, ce terme est de plus en plus utilisé pour décrire un large bassin de données, au sein duquel le schéma et les besoins relatifs à ces données ne sont pas définis jusqu'à ce qu'une requête soit effectuée.

Au sein d'une Data Warehouse, on trouve uniquement des données traitées et structurées. Dans un Data Lake, elles peuvent être structurées, semi-structurées, non-structurées ou brutes. Avant de charger une donnée dans une Data Warehouse, il est nécessaire de lui donner une forme et une structure, par exemple un modèle. Au sein d'un Data Lake, il est possible de charger des données brutes, et de leur conférer une forme et une structure uniquement quand le moment est venu d'utiliser ces données.

Le stockage dans une Data Warehouse est très cher pour les grands volumes de données, tandis qu'un Data Lake comme Hadoop est conçu pour un stockage low-cost. Deux raisons à cela. Tout d'abord, Hadoop est un logiciel open source. De fait, la licence et le support communautaire sont gratuits. De plus, Hadoop est conçu pour être installé sur du hardware bon marché.

La Warehouse est moins agile, sa configuration est figée. Il est possible de modifier sa structure, mais cela nécessite du temps. De son côté, le Lake est très agile et peut être configuré ou reconfiguré à volonté. Les modèles, les requêtes, et les applications peuvent être aisément modifiées par les développeurs et les *Data Scientists*. En revanche, la Data Warehouse est plus mature en termes de sécurité, grâce à des décennies d'existence. Malgré tout, de nombreux efforts sont déployés par l'industrie du Big Data pour sécuriser les *Data Lakes*, et le retard devrait être rapidement rattrapé. Enfin, les principaux utilisateurs du Data Lake sont des *Data Scientists*, tandis que la Data Warehouse est ouverte à tous les membres de l'entreprise.

Ces différences sont importantes à prendre en compte. Elles démontrent que le Data Lake n'est pas une simple évolution ou un remplacement de la Data Warehouse. Le Data Lake est la Data Warehouse sont optimisés pour des usages différents, et doivent être utilisés pour ce pour quoi ils ont été conçus.

Comment éviter qu'un Data Lake se transforme en Data Swamp ?

Largement adoptés par les entreprises dans le cadre d'une stratégie analytique, les *Data Lakes* se révèlent pourtant bien souvent inefficaces, en plus d'être onéreux. Afin d'éviter que votre lac de données se transforme en « marais », il est important d'éviter de commettre certaines erreurs. Voici les 5 erreurs les plus communes.

Le manque d'expérience du data lake

De nombreuses entreprises entreprennent le déploiement d'un data lake sans expérience réelle dans ce domaine. Les départements informatiques découvrent Hadoop pour la première fois, et se retrouvent rapidement désorientés par la nouveauté. Par conséquent, le déploiement est lent, l'implémentation difficile, et les objectifs sont rapidement obsolètes. Pour éviter ce problème, il est préférable de faire appel à des leaders expérimentés ou, à défaut, de contacter des experts pour leur demander conseil.

Le manque de compétences en ingénierie

Le manque d'ingénieurs qualifiés est également l'une des sources d'échec les plus courantes dans le domaine des *Data Lakes*. Il est difficile d'identifier un ingénieur qualifié pour un tel programme. La maîtrise de technologies comme Spark, Kafka et HBase est primordiale mais ne suffit pas toujours au déploiement efficace d'un

data lake. Pour minimiser ce risque, il est indispensable de recruter des ingénieurs logiciels talentueux, pour ensuite les former à la maîtrise d'Hadoop si nécessaire. Il est également possible d'investir dans une plateforme de gestion de data lake pour éviter d'avoir à utiliser Hadoop.

Un modèle d'exploitation immature

La séparation classique entre le département informatique et le reste de l'activité peut être un obstacle important, notamment pendant la phase initiale. Le déploiement réussi d'un Data Lake dépend de la collaboration étroite entre *data scientists* et *data engineers*. Les *data scientists* doivent utiliser les outils procurés par le département informatique, et les *data engineers* doivent exploiter ce qui est implémenté par les *data scientists*. Un modèle d'exploitation efficace doit être mis en place, au même titre qu'une structure de gouvernance robuste.

Une mauvaise gouvernance de données

La gouvernance de données est l'ensemble des processus assurant la gestion des ensembles de données de l'entreprise. La gouvernance garantit que les données sont dignes de confiance et que les responsables peuvent être facilement identifiés en cas de problème. Une mauvaise gouvernance est la source de nombreux échecs. La gouvernance doit être établie en amont, durant la phase initiale de l'implémentation.

Le manque de capacités techniques

Beaucoup d'entreprises sous-estiment la complexité technique des solutions data lake. Hadoop fournit plusieurs outils permettant d'implémenter une partie des capacités techniques nécessaires, mais pas toutes. Avant d'entamer l'ingestion de données, il est donc préférable de s'assurer d'avoir ces capacités à portée de main.